LA-UR-21-24972
Approved for public release; distribution is unlimited.

Title: mystic: software for autonomous discovery and design under uncertainty

Author(s): McKerns, Michael

Intended for: Report

Issued: 2021-10-26 (rev.1)

mystic: software for autonomous discovery and design under uncertainty

Michael McKerns[1,*]

[1]*The Uncertainty Quantification Foundation, Wilmington, DE 19801-6601*
[1]*Los Alamos National Laboratory, Los Alamos, NM 87545-0000*
[*]*mmckerns@uqfoundation.org*

1 Introduction

Throughout the diverse range of science and engineering applications, there is a growing desire to develop computational methods that can reliably predict the behavior of complex systems. Specifically, there is a strategic need for tools that can robustly forecast the behavior of complex physical systems, where data may be high-dimensional, noisy, or sparse, and models of the system may be time-dependent or include uncertainty. We use *mystic* [9, 11] to build tools that leverage statistical learning, physics-informed learning, and active learning in the efficient generation of reliably predictive surrogates for complex physical systems. *mystic* is a robust, proven, open-source optimization and uncertainty quantification toolkit with over a decade of use in the design and optimization of neutron instrumentation [2], solar-powered drones [5], and gas-guns [13, 15, 4], and in iterative tuning of models for Raman spectroscopy [6] and elastoplastic materials strength [13]. Recent developments have focused on automated learning of statistically robust surrogates under uncertainty, with applications in materials in extreme environments [1], nanostructures [7], materials simulations and strength models [12], and the failure of shielding under particle radiation [8]. In 2020, McKerns demonstrated active learning of optimally robust surrogates with respect to new simulated data for molecular dynamics simulations of materials mixing in warm dense matter [3], and is currently applying active learning to the automated steering of particle accelerator beams and the optimal design and control of quantum optical sensor instrumentation.

2 Methods

Estimation methods, such as interpolation and machine learning, are commonly used to build computationally inexpensive surrogates for physical models. Using an estimator to build a surrogate generally is a two-step process. First, a surrogate is fit by "training" on a subset of the existing data; then the surrogate's performance is "tested" on the remaining (holdout) data. If the surrogate's testing "score" is sufficiently good, we denote the surrogate as being *instantaneously* valid with respect to all existing data. Often, the validation of an estimator against real-world data generally is a laborious manual process that lacks a well-defined notion of solution. As a result, different teams with the same goals and data will find vastly different estimators, solutions, and notions of uncertainties. Commonly, it is assumed that if any new data is encountered, the surrogate will perform approximately as well as it did with respect to the holdout data. This assumption is often poor, especially when data is sparse or noisy, or the model is highly nonlinear, time-dependent, or includes uncertainty. Worse still, when new data eventually exposes a surrogate as a poor approximation, there is often not a well-defined process to update the surrogate. As we are interested in finding a surrogate that is *asymptotically* valid with respect to any future data, we will pose the online learning of a robust surrogate as a global optimization, where iterative updates to candidate surrogates are driven by the time-evolution of the surrogate performance with respect to new data. Fig. 1 shows an online learning procedure for generating robust predictive surrogates for complex systems. For example, we can use *mystic* to tune a neural network (ANN) to minimize the upper bound on the expected error between the predictions from the ANN and results generated by an expensive model (Fig. 2), where this entire workflow is used within `retrain(surrogate)` in Fig. 1 to ensure the surrogate is always optimally robust with respect to the data. Alternately, we can use Fig. 1 to generate a surrogate for `f(x|h)` in Fig. 2, when `f` is prohibitively expensive within the calculation of the most robust model for the existing data.

3 Examples

A common goal in areas of science and engineering is to guarantee the quality of assessments made for performance and risk in complex systems. Often, the knowledge of the system is incomplete or contains some

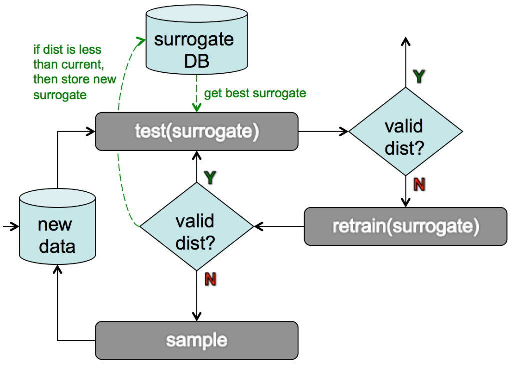

Figure 1: The current surrogate is retrieved from the surrogate DB and then tested for validity whenever new data is collected. If the surrogate is valid, it is then evaluated; otherwise, it is retrained and again checked for validity. If the surrogate is still not valid, then we use a sampler to request new data points. When a surrogate improves upon the best test score, the surrogate is stored.

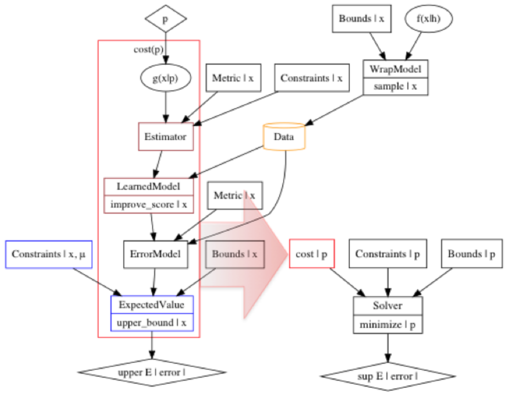

Figure 2: The inner loop trains the LearnedModel (e.g. an ANN) on Data sampled from an expensive function `f(x|h)`, and computes the upper bound on the expected misfit between `f` and the ANN. The outer loop varies the hyperparameters `p` while minimizing the expected misfit.

form of uncertainty. One promising approach to address this challenge is OUQ developed by Owhadi et al [13, 15, 10]. OUQ employs operator-like coordinate transforms to describe a system under uncertainty as a model with a set of possible outcomes formulated on a discrete probability distribution. OUQ produces a constrained global optimization, with optimization parameters composed of the probability-weighted input coordinates and a statistical objective (i.e. minimize expected misfit). OUQ requires that any knowledge of the system to be utilized in the model validation must be quantified and incorporated as constraints in the global optimization. As in Fig. 2, we leverage *mystic* to generate a machine-learned surrogate for some expensive model (i.e. a simulation or instrument), and an inverse method (like OUQ) to calculate the bounds on the expected misfit under uncertainty. *mystic* also provides a means of imposing constraining information from different sources on the search space of an optimizer. Once all available information has been encoded as constraints, the resulting information-constrained search space is used to produce an optimal estimator. This guarantees that all candidate solutions in the information-constrained search space are valid, and solutions that are most critical for design decisions are found at the bounds. In fact, the calculation of rigorous bounds under uncertainty is fundamental in the design and construction of an optimal estimator – and is, in essence, what the combination of OUQ and *mystic* provides. Surrogates generated with the above procedure can be seen in Figs. 3 and 4. Fig. 3 highlights the importance of the strategy used in sampling new points. An ensemble of optimizers is shown to effectively discover all the critical points of the Rastrigin function, while traditional sampling strategies (e.g. sampling from a Gaussian distribution) only capture the overarching trend. Fig. 4 further highlights how optimizer-driven sampling generates a higher density of samples at critical points on the surface, and facilitates the generation of a valid surrogate for a complex model.

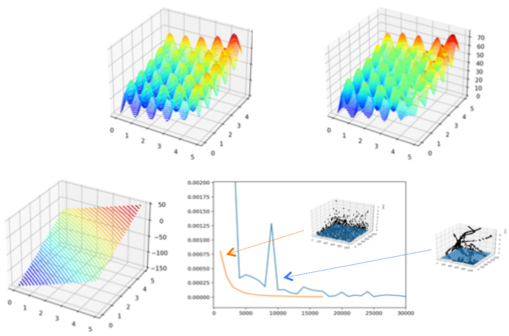

Figure 3: Test scores for actively learned surrogates of the Rastrigin function, as a function of the sampling strategy. Test scores for sampling from a Gaussian distribution (orange) converge more quickly than optimizer-directed sampling (blue); however, are not as accurate. Plots are, clockwise, the surrogate learned using Gaussian sampling, truth, and the surrogate learned using optimizer-directed sampling.

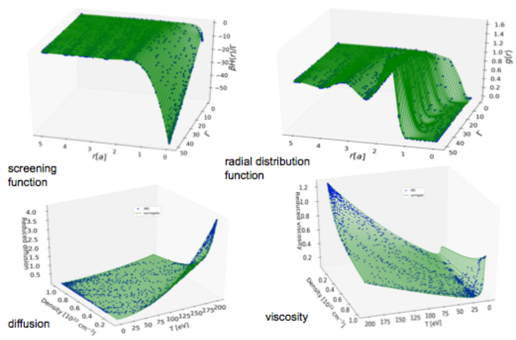

Figure 4: Robust surrogates for data generated from a one-component plasma LAMMPS [14] simulation. Surrogates for the radial distribution function and screening function were simultaneously learned, similarly for the self-diffusion and viscosity. The blue dots highlight how optimizer-driven sampling generates a higher density of samples at critical points on the surface.

4 Acknowledgments

Research presented in this article was supported by Los Alamos National Laboratory under the Laboratory Directed Research and Development program (project numbers 20190005DR and 20200410DI), by the Department of Energy Advanced Simulation and Computing under the Beyond Moore's Law Program, and by the Uncertainty Quantification Foundation under the Statistical Learning program. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). The Uncertainty Quantification Foundation is a nonprofit dedicated to the advancement of predictive science through research, education, and the development and dissemination of advanced technologies.

References

[1] J. Belak, D. Orikowski, S. Applegate, H. Owhadi, and M. McKerns. Quantifying model uncertainty, 2012. LLNL-PRES-585774.

[2] C.M. Biwer, S.C. Vogel, M.M. McKerns, and J.P. Ahrens. Spotlight: Distributed-computing for rietvled analyses using an ensemble of local optimizers, 2019. http://github.com/lanl/spotlight.

[3] A. Diaw, M. McKerns, I. Sagert, L.G. Stanton, and M.S. Murillo. Accurate surrogates of microscopic data with directed sampling, 2020. LA-UR-20-24947.

[4] P.-H. T. Kamga, B. Li, M. McKerns, L. H. Nguyen, M. Ortiz, H. Owhadi, and T. J. Sullivan. Optimal uncertainty quantification with model uncertainty and legacy data. *Journal of the Mechanics and Physics of Solids*, 72:1–19, 2014.

[5] F. Khoshnoud, I. I. Esat, C. W. de Silva, M. M. McKerns, and H. Owhadi. Self-Powered Dynamic Systems in the Framework of Optimal Uncertainty Quantification. *Journal of Dynamic Systems, Measurement, and Control*, 139(9). doi: 10.1115/1.4036367.

[6] C. W. Li, M. M. McKerns, and B. Fultz. A raman spectrometry study of phonon anharmonicity of zirconia at elevated temperatures. *Journal of the American Ceramic Society*, 94(1):224–229, 2011. doi: 10.1111/j.1551-2916.2010.04057.x.

[7] M. McKerns. Is Automated Materials Design and Discovery Possible? In T. Lookman, S. Eidenbenz, F. Alexander, and C. Barnes, editors, *Materials Discovery and Design*, volume 280 of *Springer Series in Materials Science*, pages 15–58. Springer, Cham., 2018. https://doi.org/10.1007/978-3-319-99465-9_2.

[8] M. McKerns, L. Roth, N. Iyengar, and D. Lamm. Rigorous bounds on the failure of shielding due to helium-ion radiation. in preparation.

[9] M. McKerns, P. Hung, and M. Aivazis. mystic: highly-constrained non-convex optimization and UQ, 2009. http://pypi.python.org/pypi/mystic.

[10] M. McKerns, H. Owhadi, C. Scovel, T. J. Sullivan, and M. Ortiz. The optimal uncertainty algorithm in the mystic framework. Caltech CACR Technical Report, August 2010. http://arxiv.org/pdf/1202.1055v1.

[11] M. McKerns, L. Strand, T. J. Sullivan, A. Fang, and M. Aivazis. Building a framework for predictive science. In *Proceedings of the 10th Python in Science Conference*, pages 67–78, 2011. http://arxiv.org/pdf/1202.1056.

[12] M. McKerns, F.J. Alexander, K.S. Hickmann, T.J. Sullivan, and D.E. Vaughn. Optimal Bounds on Nonlinear Partial Differential Equations in Model Certification, Validation, and Experiment Design, 2020. https://arxiv.org/abs/2009.06626.

[13] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal Uncertainty Quantification. *SIAM Rev.*, 55(2):271–345, 2013. doi: 10.1137/10080782X.

[14] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J Comp Phys*, 117:1–19, 1995.

[15] T. J. Sullivan, M. McKerns, D. Meyer, F. Theil, H. Owhadi, and M. Ortiz. Optimal uncertainty quantification for legacy data observations of Lipschitz functions. *ESAIM Math. Model. Numer. Anal.*, 47 (6):1657–1689, 2013. doi: 10.1051/m2an/2013083.